

On Varieties of Doubly Robust Estimators Under Missingness Not at Random With a Shadow Variable

BY WANG MIAO

*Beijing International Center for Mathematical Research, Peking University,
Beijing 100871, P.R.C.
mwfy@pku.edu.cn*

AND ERIC TCHETGEN TCHETGEN

*Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.
etchetge@hsph.harvard.edu*

SUMMARY

Suppose we are interested in the mean of an outcome variable missing not at random. Suppose however that one has available a fully observed shadow variable, which is associated with the outcome but independent of the missingness process conditional on covariates and the possibly unobserved outcome. Such a variable may be a proxy or a mismeasured version of the outcome available for all individuals. We have previously established necessary and sufficient conditions for identification of the full data law in such a setting, and have described semiparametric estimators including a doubly robust estimator of the outcome mean. Here, we propose two alternative doubly robust estimators for the outcome mean, which may be viewed as extensions of analogous methods under missingness at random, but enjoy different properties. We assess correctness of the required working models via straightforward goodness-of-fit tests.

Some key words: Doubly robust estimation; Missingness not at random; Shadow variable.

1. INTRODUCTION

Doubly robust methods are designed to mitigate estimation bias due to model misspecification in observational studies and imperfect experiments. Such methods have grown in popularity in recent years for estimation with missing data and other forms of coarsening (Robins et al., 1994; Scharfstein et al., 1999; Van der Laan & Robins, 2003; Bang & Robins, 2005; Tsiatis, 2006). There exist various constructions of doubly robust estimators for the mean of an outcome that is missing at random; see Kang & Schafer (2007). In contrast, for data missing not at random, difficulty of identification undermines one's ability to obtain accurate inferences, and doubly robust estimation is far more challenging. Identification of a full data model means that, the parameters indexing the model are uniquely determined by the observed data, i.e., the data that are actually observed on the individuals. Statistical inference based on non-identifiable models may be misleading and of limited interest in practice; see Miao et al. (2015). Under missingness at random, the full data law, i.e., the joint distribution of all variables of interest, is nonparametrically identified from the observed data. However, under missingness not at random, identification is no longer possible without further restrictions on the missingness process. Although no general identification results are available for data missing not at random, one may identify the full data

law under specific assumptions. Building on earlier work by D’Haultfoeuille (2010), Wang et al. (2014) and Zhao & Shao (2014), Miao et al. (2015) used a fully observed shadow variable to establish a general identification framework for data missing not at random. Such a variable is associated with the outcome conditional on covariates, but independent of the missingness conditional on covariates and the outcome (Kott, 2014); it may be available in many empirical studies, where a fully observed proxy or a mismeasured version of the outcome is available. For example, in a study of mental health of children in Connecticut (Zahner et al., 1992; Ibrahim et al., 2001), researchers were interested in evaluating the prevalence of students with abnormal psychopathological status based on their teacher’s assessment, which was subject to missingness. A separate parent report available for all children in the study, is a proxy for the teacher’s assessment, but is unlikely to be related to the teacher’s response rate conditional on covariates and her assessment of the student; in this case the parental assessment constitutes a valid shadow variable. Other examples can be found in Miao et al. (2015) and Wang et al. (2014).

Throughout, we let Y denote the outcome, R is its missingness indicator with $R = 1$ if Y is observed, otherwise $R = 0$, and let X denote fully observed covariates. Suppose that one has also fully observed a shadow variable Z that satisfies

Assumption 1. (i) $Z \not\perp Y \mid X$; (ii) $Z \perp R \mid (Y, X)$.

Assumption 1 formalizes the idea that, the shadow variable only affects the missingness through its association with the outcome. We provide a directed acyclic graph in the Supplementary Material that can help to understand the assumption. The shadow variable introduces additional conditional independence conditions, which impose further restrictions on the missingness process, and thus provides better opportunity for identification despite the fact that data may be missing not at random. Miao et al. (2015) presented a brief review of such problems, and gave necessary and sufficient conditions as well as sufficient conditions for identification with a shadow variable. In particular, if the outcome is binary, the full data law is identifiable with a binary shadow variable. But for a continuous outcome, a binary shadow variable does not impose enough restrictions to identify the full data law; see the Supplement Material for a counterexample. Identification for a continuous outcome requires at least one continuous shadow variable, but even then, additional conditions are needed. We consider a location-scale model for the density function:

$$f(y \mid x, z, r) = \frac{1}{\sigma_r(z, x)} f_r \left\{ \frac{y - \mu_r(z, x)}{\sigma_r(z, x)} \right\}, \quad r = 0, 1, \quad (1)$$

with unrestricted functions μ_r and σ_r , and density functions f_r . Under certain regularity conditions summarized in the Appendix, we have previously proved identification of the full data law if either $f(y \mid x, z, r = 1)$ or $f(y \mid x, z, r = 0)$ follows model (1), even if the missingness process is unrestricted (Miao et al., 2015). Aside for Assumption 1, model (1) includes many commonly-used models, for instance, Gaussian models, and thus essentially demonstrates that lack of identification is not an issue in many familiar situations. However, one cannot understate the central role of the shadow variable for identification. Without such a variable, identification is no longer guaranteed for model (1), even if one were to assume a parametric missingness model. For additional and extensive discussion about identification under missingness not at random with a shadow variable, see Miao et al. (2015) and Wang et al. (2014).

With models satisfying the corresponding identification conditions, previous authors have developed several non-doubly robust estimators. Among them, inverse probability weighted estimation (Wang et al., 2014) and pseudo-likelihood estimation (Zhao & Shao, 2014) are sensitive to model misspecification; and nonparametric estimation (D’Haultfoeuille, 2010) requires an unrealistic large sample size for reasonable performance when the covariate dimension is moderate

to large. In contrast, a doubly robust approach remains consistent and asymptotically normal under partial misspecification. Specifically, Miao et al. (2015) developed a doubly robust estimator based on a three-part model for the full data: a model for the joint distribution of the outcome and the shadow variable in complete cases; a model for the propensity score evaluated at a reference value of the outcome; and a log odds ratio model encoding the association of the outcome and the missingness process. Under correct specification of the log odds ratio model, the doubly robust estimator is consistent if either of the other two models is correct, but not necessarily both. However, the construction of a doubly robust estimator is not unique. In this paper, we develop two alternative doubly robust estimators of the outcome mean that enjoy different properties, and we compare them both in theory and via simulations reported in the Supplementary Material.

2. DOUBLY ROBUST ESTIMATORS

Under Assumption 1, we factorize the conditional density function of (Z, Y, R) given X as

$$f(z, y, r | x) = c(x) \exp\{(1 - r)\text{OR}(y | x)\} \text{pr}(r | y = 0, x) f(z, y | r = 1, x), \quad (2)$$

where $c(x) = \text{pr}(r = 1 | x) / \text{pr}(r = 1 | y = 0, x)$; $\text{pr}(r = 1 | y = 0, x)$ is the response probability evaluated at the reference level $y = 0$, and is referred to as the baseline propensity score; $f(z, y | r = 1, x)$ is the joint density function of (Z, Y) conditional on X among the complete cases, i.e., the subset with $r = 1$, and is referred to as the baseline outcome density;

$$\text{OR}(y | x) = \log \frac{\text{pr}(r = 0 | y, x) \text{pr}(r = 1 | y = 0, x)}{\text{pr}(r = 0 | y = 0, x) \text{pr}(r = 1 | y, x)},$$

is the log of the conditional odds ratio function relating Y and R given X with $E[\exp\{\text{OR}(y | x)\} | r = 1, x] < \infty$ and $\text{OR}(y = 0 | x) = 0$. For a continuous outcome, we require that $f(z, y | r = 1, x)$ satisfies model (1) to guarantee identification. For estimation, we specify separate parametric models $\text{pr}(r = 1 | y = 0, x; \alpha)$, $f(z, y | r = 1, x; \beta)$, and $\text{OR}(y | x; \gamma)$. We suppose throughout that $\text{OR}(y | x; \gamma)$ is correctly specified, which can be achieved by specifying a relatively flexible model, or following the approach suggested by Higgins et al. (2008) if information on the reasons for missingness are available. From (2), we have the following identities:

$$\text{pr}(r = 1 | y, x) = \frac{\text{pr}(r = 1 | y = 0, x)}{\text{pr}(r = 1 | y = 0, x) + \exp\{\text{OR}(y | x)\} \text{pr}(r = 0 | y = 0, x)}, \quad (3)$$

$$f(z, y | r = 0, x) = \frac{\exp\{\text{OR}(y | x)\}}{E[\exp\{\text{OR}(y | x)\} | r = 1, x]} f(z, y | r = 1, x), \quad (4)$$

$$E(y | r = 0, x) = \frac{E[\exp\{\text{OR}(y | x)\} y | r = 1, x]}{E[\exp\{\text{OR}(y | x)\} | r = 1, x]}. \quad (5)$$

The propensity score, and its reciprocal, i.e., the inverse probability weight function $W(x, y; \alpha, \gamma) = 1 / \text{pr}(r = 1 | x, y; \alpha, \gamma)$, are determined by the baseline propensity score model $\text{pr}(r = 1 | x, y = 0; \alpha)$ and the log odds ratio model $\text{OR}(y | x; \gamma)$ as in (3); the conditional outcome mean among the incomplete cases $E(y | r = 0, x; \beta, \gamma)$ is determined by the baseline outcome model and the log odds ratio model as in (5).

Estimation of β only involves the complete cases. Let \hat{E} denote the empirical mean, we solve

$$\hat{E}\{rS(z, y, x; \hat{\beta})\} = 0, \quad (6)$$

with score function $S(z, y, x; \beta) = \partial \log\{P(z, y | r = 1, x; \beta)\} / \partial \beta$. Estimation of $\hat{\alpha}$ and $\hat{\gamma}$ is motivated from a classic estimating equation following the fact that the respective weighted

mean of any vector functions $G(x, y)$ and $H(x)$ among the complete cases equals their population mean: $\widehat{E}[\{W(x, y; \widehat{\alpha}, \widehat{\gamma})r - 1\}\{G(x, y)^T, H(x)^T\}^T] = 0$, where $G(x, y)$ and $H(x)$ are user-specified vector functions of dimension equal to that of γ and α , respectively, and satisfy $E[\partial W(x, y; \alpha, \gamma)r / \partial(\alpha, \gamma)\{G(x, z)^T, H(x)^T\}]$ is nonsingular for all (α, γ) . For example, if $\text{pr}(r = 1 | y, x; \alpha, \gamma)$ follows a logistic model and thus $W(x, y; \alpha, \gamma) = 1 + \exp\{-(1, x^T)\alpha - \gamma y\}$, we may naturally choose $G(x, y) = y$ and $H(x) = (1, x^T)^T$. Because y is missing for $r = 0$, the classic estimating equation is not feasible. However, Assumption 1 allows us to replace y with the shadow variable z and to replace $G(x, y)$ with $G(x, z)$. To further derive doubly robust estimators, we incorporate the baseline outcome model into the estimating equation for (α, γ) . Let $G_1(x, z; \beta, \gamma) = G(x, z) - E\{G(x, z) | r = 0, x; \beta, \gamma\}$, we solve

$$\widehat{E}[\{W(x, y; \widehat{\alpha}, \widehat{\gamma})r - 1\}\{G_1(x, z; \widehat{\beta}, \widehat{\gamma})^T, H(x)^T\}^T] = 0, \quad (7)$$

with $G(x, z)$ and $H(x)$ such that $E[\partial W(x, y; \alpha, \gamma)r / \partial(\alpha, \gamma)\{G_1(x, z; \beta, \gamma)^T, H(x)^T\}]$ is nonsingular for all (α, β, γ) . The shadow variable Z is used as a proxy of Y , thus, a choice of Z that is highly correlated with Y is desirable for the purpose of efficiency maximization.

Using $(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma})$ obtained from equations (6) and (7), we construct three different estimators for the outcome mean that are consistent if either the baseline outcome model or the baseline propensity score model is correctly specified, together with the log odds ratio model.

A regression estimator with residual bias correction was previously described by Miao et al. (2015). We use the weighted residual to correct the bias of the conditional mean among incomplete cases. Let $M_0(x; \widehat{\beta}, \widehat{\gamma}) = E(y | r = 0, x; \widehat{\beta}, \widehat{\gamma})$, the estimator is

$$\widehat{\mu}_1 = \widehat{E}[W(x, y; \widehat{\alpha}, \widehat{\gamma})r\{y - M_0(x; \widehat{\beta}, \widehat{\gamma})\} + M_0(x; \widehat{\beta}, \widehat{\gamma})].$$

A Horvitz–Thompson estimator with extended weights employs an extended baseline propensity score model and an extended weight function. The extended baseline propensity score model with unknown parameter ϕ satisfies $\text{pr}_{\text{ext}}(r = 1 | y = 0, x; \phi) = \text{pr}(r = 1 | y = 0, x; \widehat{\alpha})$ only at $\phi = 0$. For example, we can specify

$$\text{pr}_{\text{ext}}(r = 1 | y = 0, x; \phi) = \frac{\text{pr}(r = 1 | y = 0, x; \widehat{\alpha})}{\text{pr}(r = 1 | y = 0, x; \widehat{\alpha}) + \exp\{\phi g(x)\}\text{pr}(r = 0 | y = 0, x; \widehat{\alpha})},$$

with user-specified scalar function $g(x)$. The extended weight function $W_{\text{ext}}(x, y; \phi)$, and its reciprocal is determined as in (3) with $OR(y|x)$ and $\text{pr}(r = 1 | y = 0, x)$ replaced by $OR(y | x; \widehat{\gamma})$ and $\text{pr}_{\text{ext}}(r = 1 | y = 0, x; \phi)$ respectively. We estimate ϕ by solving

$$\widehat{E}[\{W_{\text{ext}}(x, y; \widehat{\phi})r - 1\}\{M_0(x; \widehat{\beta}, \widehat{\gamma}) - \widehat{\mu}_{\text{reg}}\}] = 0, \quad (8)$$

with previously obtained $(\widehat{\beta}, \widehat{\gamma})$ and $\widehat{\mu}_{\text{reg}} = \widehat{E}\{(1 - r)M_0(x; \widehat{\beta}, \widehat{\gamma}) + ry\}$. The Horvitz–Thompson estimator with extended weights is

$$\widehat{\mu}_2 = \widehat{E}\left\{\frac{W_{\text{ext}}(x, y; \widehat{\phi})r}{\widehat{E}\{W_{\text{ext}}(x, y; \widehat{\phi})r\}}y\right\}.$$

A regression estimator with an extended outcome model involves an extended outcome model $M_{0\text{ext}}(x; \psi)$ with parameter ψ satisfying $M_{0\text{ext}}(x; \psi) = M_0(x; \widehat{\beta}, \widehat{\gamma})$ only at $\psi = 0$. If $M_0(x; \widehat{\beta}, \widehat{\gamma}) = \lambda\{Q(x; \widehat{\beta}, \widehat{\gamma})\}$ for some inverse link λ and some function Q , we can specify $M_{0\text{ext}}(x; \psi) = \lambda\{Q(x; \widehat{\beta}, \widehat{\gamma}) + \psi q(x)\}$ with a scalar function $q(x)$. We estimate ψ by solving

$$\widehat{E}[\{W(x, y; \widehat{\alpha}, \widehat{\gamma}) - 1\}r\{y - M_{0\text{ext}}(x; \widehat{\psi})\}] = 0, \quad (9)$$

with previously obtained $(\hat{\alpha}, \hat{\gamma})$. The regression estimator with an extended outcome model is

$$\hat{\mu}_3 = \hat{E}\{(1 - r)M_{0\text{ext}}(x; \hat{\psi}) + ry\}.$$

The estimators $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\mu}_3$ may have very different characteristics, although, all three estimators are doubly robust.

THEOREM 1. *Under Assumption 1, if the log odds ratio model $\text{OR}(y | x; \gamma)$ is correct, and the probability limit of equations (6), (7), (8) and (9) has a unique solution, then $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\mu}_3$ are consistent if either $f(z, y | r = 1, x; \beta)$ or $\text{pr}(r = 1 | y = 0, x; \alpha)$ is correctly specified.*

The extended models not only provide double robustness, but also provide a strategy to check if the working models are correct. We prove in the Appendix that if the baseline propensity score model is correct, $\hat{\phi}$ converges to 0 in probability; and if the baseline outcome model is correct, $\hat{\psi}$ converges to 0 in probability. Therefore, one may use this property to assess whether the working models are correctly specified by checking whether $\hat{\phi}$ and $\hat{\psi}$ are within sampling variability of zero, respectively. However, one should acknowledge that the space of possible departures from the assumed model may be prohibitively large relative to the proposed test so that the resulting goodness-of-fit test will generally have good power against certain alternatives but not in all possible directions away from the specified working model. We explore the power of the proposed goodness-of-fit test via a simulation study in the Supplementary Material.

All three doubly robust estimators rely on a correct log odds ratio model, since inference about the law of Y requires an accurate evaluation of the dependence between the missingness process and the outcome, which is captured by the log odds ratio model $\text{OR}(y | x; \gamma)$. To the best of our knowledge, with the exception of Miao et al. (2015), previous doubly robust estimators have assumed that this log odds ratio is known, either to equal the null value of 0 under missingness at random (Bang & Robins, 2005; Tsiatis, 2006; Van der Laan & Robins, 2003), or to be of a known functional form with no unknown parameters (Vansteelandt et al., 2007; Robins et al., 2008). We have relaxed these more stringent assumptions.

3. RELATION TO PREVIOUS DOUBLY ROBUST ESTIMATORS AND COMPARISONS

Previous doubly robust estimators under missingness at random can be viewed as special cases of our estimators. Under missingness at random, $\text{OR}(y | x) = 0$, $\text{pr}(r = 1 | x, y = 0) = \text{pr}(r = 1 | x)$, the inverse probability weight function $W(x; \alpha) = 1/\text{pr}(r = 1 | x; \alpha)$ does not vary with y , and the conditional mean among the population $M(x; \beta)$ equals that among the incomplete cases $M_0(x; \beta, \gamma)$. The estimator $\hat{\mu}'_1 = \hat{E}[W(x; \hat{\alpha})r\{y - M(x; \hat{\beta})\} + M(x; \hat{\beta})]$ of Kang & Schafer (2007) is a special case of the regression estimator with residual bias correction; the estimator $\hat{\mu}'_2 = \hat{E}[W_{\text{ext}}(x; \hat{\phi})r/\hat{E}\{W_{\text{ext}}(x; \hat{\phi})r\}y]$ proposed by Robins et al. (2007), with an extended logistic propensity score model $\text{logit pr}_{\text{ext}}(r = 1 | x; \phi) = (1, x^T)\hat{\alpha} + \phi g(x)$, is a special case of the Horvitz–Thompson estimator with extended weights; the estimator $\hat{\mu}'_3 = \hat{E}\{M_{\text{ext}}(x; \hat{\psi})\}$ proposed by Robins et al. (2007), with an extended outcome model $M_{\text{ext}}(x; \hat{\psi})$ satisfying $\hat{E}[W(x; \hat{\alpha})r\{y - M_{\text{ext}}(x; \hat{\psi})\}] = 0$ and $\hat{E}[r\{y - M_{\text{ext}}(x; \hat{\psi})\}] = 0$, is a special case of the regression estimator with an extended outcome model.

The three proposed doubly robust estimators enjoy some of the properties of their missingness at random analogs. The estimator $\hat{\mu}_2$ is a convex combination of the observed outcome values. It satisfies the boundedness property (Robins et al., 2007) that the estimator falls in the parameter space for the outcome mean almost surely. Such estimators are preferred when the inverse probability weights are highly variable, because they rule out estimates outside the sample space.

Boundedness is not guaranteed for $\hat{\mu}_1$. If the range of $M_{0\text{ext}}(x; \psi)$ is contained in the sample space of the outcome, $\hat{\mu}_3$ also satisfies the boundedness condition, but this does not hold in general. For example, if the outcome is continuous, and $M_{0\text{ext}}(x; \psi) = M_0(x; \hat{\beta}, \hat{\gamma}) + \psi$, the range of $\hat{\mu}_3$ may be outside the sample space of the outcome mean.

The three proposed estimators offer certain improvements in term of bias when both models are misspecified. The asymptotic bias of $\hat{\mu}_1$ can be written as

$$\text{Bias}_1 = E[\{W(x, y; \alpha^*, \gamma^*)r - 1\}\{y - M_0(x; \beta^*, \gamma^*)\}],$$

and the asymptotic bias of $\hat{\mu}_3$ has the same form with $M_0(x; \beta, \gamma^*)$ replaced by $M_{0\text{ext}}(x; \psi^*)$, with probability limits $(\alpha^*, \beta^*, \gamma^*, \psi^*)$ of the corresponding estimators. The bias is driven by the degree of misspecification of both the weight function and the conditional mean among the incomplete cases. As pointed out by Robins et al. (2007) and Vermeulen & Vansteelandt (2014), without further restrictions on the inverse probability weights, Bias_1 gets inflated in regions with large weights. However, if the components of $H(x)$ in equation (7) include a constant function, then $E\{W(x, y; \alpha^*, \gamma^*)r\} = 1$, which restricts the amount of variability of the inverse probability weights. Thus, Bias_1 does not explode with large weights.

In simulation studies, we found that the three doubly robust estimators approximate the true outcome mean if either of the baseline models is correct, but they are biased if neither baseline model is correct. For the case with moderately variable weights, the relative magnitude of the bias depends on the specific data generating process, but for the case with highly-variable weights, the Horvitz–Thompson estimator with extended weights has smaller bias. If the baseline outcome model is correct, the parameter of the extended outcome model, $\hat{\psi}$ is close to 0; and if the baseline propensity score model is correct, the parameter of the extended weight model, $\hat{\phi}$ is close to 0. We also perform formal tests of the null hypotheses $\mathbb{H}_0 : \phi = 0$ and $\mathbb{H}_0 : \psi = 0$ respectively under level 0.05. The results show an empirical type I error approximating 0.05 if the required baseline propensity score model or baseline outcome model is correct, respectively (i.e., the true value of ϕ and ψ equals 0 respectively). Such tests have good power in moderate samples if the required model is incorrect, respectively. We recommend the proposed hypothesis tests to check for severe misspecification of the baseline models in practice.

4. DISCUSSION

Extensions of the doubly robust methods described in this work to other functionals, such as a parameter δ solving a full data estimating equation $E\{U(z, y, x; \delta)\} = 0$, can be achieved by replacing Y with U wherever Y occurs in the estimating equations and solving the doubly robust estimating equation for the parameter of interest. The methods also have potential application in related areas, such as longitudinal data analysis and causal inference.

ACKNOWLEDGEMENT

The work is partially supported by the China Scholarship Council and the National Institute of Health. The authors are grateful to the referees and the editor for their helpful comments.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proof of a lemma, a counterexample to identification with a continuous outcome, a graph model for the shadow variable, and simulation studies.

APPENDIX

Proof of Theorem 1

We need the following lemma, which we prove in the Supplementary Material.

LEMMA A1. *Under Assumption 1, suppose that the log odds ratio model is correct, and that the probability limit of equations (6) and (7) has a unique solution. For any square integrable vector function $D(z, y, x)$, scalar function $V(x)$, and $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ solving equations (6) and (7),*

- (i) *if $\text{pr}(r = 1 \mid y = 0, x; \alpha)$ is correct, then $\hat{E}[\{W(x, y; \hat{\alpha}, \hat{\gamma})r - 1\}D(z, y, x)]$ converges to 0 in probability;*
- (ii) *if $f(z, y \mid r = 1, x; \beta)$ is correct, then $\hat{E}[r \exp\{\text{OR}(y \mid x; \hat{\gamma})\}V(x)\{D(z, y, x) - E[D(z, y, x) \mid r = 0, x; \hat{\beta}, \hat{\gamma}]\}]$ converges to 0 in probability;*
- (iii) *if either of the baseline models is correct, then $\hat{E}[\{W(x, y; \hat{\alpha}, \hat{\gamma})r - 1\}\{D(z, y, x) - E[D(z, y, x) \mid r = 0, x; \hat{\beta}, \hat{\gamma}]\}]$ converges to 0 in probability.*

Proof of Theorem 1. Suppose that the log odds ratio model is correctly specified, and that the probability limit of the estimating equations has a unique solution.

1. Double robustness of $\hat{\mu}_1$. If either of the baseline models is correct, from (iii) of Lemma 1, $\hat{E}[\{W(x, y; \hat{\alpha}, \hat{\gamma})r - 1\}\{y - E(y \mid r = 0, x; \hat{\beta}, \hat{\gamma})\}]$ converges to 0, therefore $\hat{E}[W(x, y; \hat{\alpha}, \hat{\gamma})r\{y - M_0(x; \hat{\beta}, \hat{\gamma})\} + M_0(x; \hat{\beta}, \hat{\gamma})]$ converges to the true outcome mean.
2. Double robustness of $\hat{\mu}_2$. From (i) of Lemma 1, if the baseline propensity score model is correct, $\hat{E}[\{W_{\text{ext}}(x, y; \phi = 0)r - 1\}\{M_0(x; \hat{\beta}, \hat{\gamma}) - \hat{\mu}_{\text{reg}}\}] = \hat{E}[\{W(x, y; \hat{\alpha}, \hat{\gamma})r - 1\}\{M_0(x; \hat{\beta}, \hat{\gamma}) - \hat{\mu}_{\text{reg}}\}]$ converges to 0, i.e., $\phi = 0$ is a solution of the probability limit of equation (8). Thus, the solution of equation (8) $\hat{\phi}$ converges to 0, and $\lim_{n \rightarrow +\infty} \hat{E}\{W_{\text{ext}}(x, y; \hat{\phi})r\} = 1$, $\lim_{n \rightarrow +\infty} \hat{E}\{W_{\text{ext}}(x, y; \hat{\phi})ry\} = \lim_{n \rightarrow +\infty} \hat{E}\{W(x, y; \hat{\alpha}, \hat{\gamma})ry\} = E(Y)$. If the baseline outcome model is correct, $\hat{E}[(1 - r)\{y - M_0(x; \hat{\beta}, \hat{\gamma})\}]$ converges to 0; $\hat{\mu}_{\text{reg}} = \hat{E}[(1 - r)M_0(x; \hat{\beta}, \hat{\gamma}) + ry]$ converges to the true outcome mean; and $\hat{E}(y - \hat{\mu}_{\text{reg}})$ converges to 0. By definition of the extended weight function, $\{W_{\text{ext}}(x, y; \hat{\phi}) - 1\}r = r \exp\{\text{OR}(y \mid x; \hat{\gamma})\}V(x)$ with $V(x) = \text{pr}_{\text{ext}}(r = 0 \mid y = 0, x; \hat{\phi})/\text{pr}_{\text{ext}}(r = 1 \mid y = 0, x; \hat{\phi})$. From (ii) of Lemma 1, $\hat{E}[\{W_{\text{ext}}(x, y; \hat{\phi}) - 1\}r\{y - M_0(x; \hat{\beta}, \hat{\gamma})\}]$ converges to 0. Thus, $\hat{E}[\{W_{\text{ext}}(x, y; \hat{\phi})r - 1\}\{y - M_0(x; \hat{\beta}, \hat{\gamma})\}]$ converges to 0, and

$$\begin{aligned} \hat{\mu}_2 &= 1/\hat{E}\{W_{\text{ext}}(x, y; \hat{\phi})r\} \cdot \hat{E}[\{W_{\text{ext}}(x, y; \hat{\phi})r - 1\}\{y - M_0(x; \hat{\beta}, \hat{\gamma})\}] \\ &\quad + 1/\hat{E}\{W_{\text{ext}}(x, y; \hat{\phi})r\} \cdot \hat{E}[\{W_{\text{ext}}(x, y; \hat{\phi})r - 1\}\{M_0(x; \hat{\beta}, \hat{\gamma}) - \hat{\mu}_{\text{reg}}\}] \\ &\quad + 1/\hat{E}\{W_{\text{ext}}(x, y; \hat{\phi})r\} \cdot \hat{E}(y - \hat{\mu}_{\text{reg}}) + \hat{\mu}_{\text{reg}} \end{aligned} \quad \square$$

converges to the true outcome mean in probability.

3. Double robustness of $\hat{\mu}_3$. If $\text{pr}(r = 1 \mid x, y = 0; \alpha)$ is correct, from (i) of Lemma 1, $\hat{E}[\{W(x, y; \hat{\alpha}, \hat{\gamma})r - 1\}\{y - M_{0\text{ext}}(x; \hat{\psi})\}]$ converges to 0. Note equation (9), we have that $\hat{E}[(1 - r)\{y - M_{0\text{ext}}(x; \hat{\psi})\}]$ converges to 0. Thus, $\hat{\mu}_3 = \hat{E}\{(1 - r)M_{0\text{ext}}(x; \hat{\psi}) + ry\}$ converges to the true outcome mean. If $f(z, y \mid r = 1, x; \beta)$ is correct, then $\hat{E}[(1 - r)\{y - M_0(x; \hat{\beta}, \hat{\gamma})\}]$ converges to 0. Since $\{W(x, y; \hat{\alpha}, \hat{\gamma}) - 1\}r = r \exp\{\text{OR}(y \mid x; \hat{\gamma})\}V(x)$ with $V(x) = \text{pr}(r = 0 \mid y = 0, x; \hat{\alpha})/\text{pr}(r = 1 \mid y = 0, x; \hat{\alpha})$, from (ii) of Lemma 1, $\hat{E}[\{W(x, y; \hat{\alpha}, \hat{\gamma}) - 1\}r\{y - M_{0\text{ext}}(x; \hat{\psi})\}] = \hat{E}[\{W(x, y; \hat{\alpha}, \hat{\gamma}) - 1\}r\{y - M_0(x; \hat{\beta}, \hat{\gamma})\}]$ converges to 0. That is, $\hat{\psi} = 0$ is a solution of the probability limit of equation (9). Thus, the solution of equation (9), $\hat{\psi}$ converges to 0, and $\lim_{n \rightarrow +\infty} \hat{E}\{(1 - r)M_{0\text{ext}}(x; \hat{\psi}) + ry\} = \lim_{n \rightarrow +\infty} \hat{E}\{(1 - r)M_0(x; \hat{\beta}, \hat{\gamma}) + ry\} = E(Y)$.

Regularity conditions for model (1)

The full data law is identifiable if either $f(y \mid z, x, r = 1)$ or $f(y \mid z, x, r = 0)$ follows the location-scale model (1), and the corresponding density function $f_{r=1}$ or $f_{r=0}$ satisfies the following conditions:

- (a) the characteristic function $\varphi(t)$ of the density function $f(v)$ satisfies $0 < |\varphi(t)| < C \exp(-\delta|t|)$ for $t \in \mathbb{R}$ and some constants $C, \delta > 0$;
- (b) conditional on x , $\mu(z, x)$, $\sigma(z, x)$ are continuously differentiable and integrable with respect to z ; $f(v)$ is continuously differentiable, and $\int_{-\infty}^{+\infty} |v \cdot \partial f(v)/\partial v|^2 dv$ is finite;
- (c) there exist some linear one-to-one mapping $M : f\{(v - a)/b\} \mapsto h(t, a, b)$ and some value $-\infty \leq t_0 \leq +\infty$ such that $\lim_{t \rightarrow t_0} h(t, a, b)/h(t, a', b')$ either equals zero or infinity for any $a, a' \in \mathbb{R}$, $b, b' > 0$ with $(a, b) \neq (a', b')$.

Many commonly-used models satisfy conditions (a)-(c), for example, the Gaussian models with f the standard normal density function, M the inverse Laplace transform, $h(t, a, b)$ the moment-generating function of a normal density function with mean a and variance b^2 , and $t_0 = +\infty$.

REFERENCES

- BANG, H. & ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.
- D’HAULTFOEUILLE, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics* **154**, 1–15.
- HIGGINS, J. P., WHITE, I. R. & WOOD, A. M. (2008). Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical Trials* **5**, 225–239.
- IBRAHIM, J. G., LIPSITZ, S. R. & HORTON, N. (2001). Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**, 361–373.
- KANG, J. D. & SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539.
- KOTT, P. (2014). Calibration weighting when model and calibration variables can differ. In *Contributions to Sampling Statistics*, Contributions to Statistics. Springer International Publishing, pp. 1–18.
- MIAO, W., DING, P. & GENG, Z. (2015). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **accepted**.
- MIAO, W., TCHETGEN TCHETGEN, E. & GENG, Z. (2015). Identification and doubly robust estimation of data missing not at random with a shadow variable. *ArXiv:1509.02556*.
- ROBINS, J., LI, L., TCHETGEN TCHETGEN, E., VAN DER VAART, A. et al. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, vol. 2. Institute of Mathematical Statistics, pp. 335–421.
- ROBINS, J., SUED, M., LEI-GOMEZ, Q. & ROTNITZKY, A. (2007). Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science* **22**, 544–559.
- ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.
- SCHARFSTEIN, D. O., ROTNITZKY, A. & ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.
- TSIATIS, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
- VAN DER LAAN, M. J. & ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
- VANSTEELANDT, S., ROTNITZKY, A. & ROBINS, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94**, 841–860.
- VERMEULEN, K. & VANSTEELANDT, S. (2014). Biased-reduced doubly robust estimation. *Journal of the American Statistical Association* **accepted**.
- WANG, S., SHAO, J. & KIM, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24**, 1097–1116.
- ZAHNER, G. E., PAWELKIEWICZ, W., DEFRANCESCO, J. J. & ADNOPOZ, J. (1992). Children’s mental health service needs and utilization patterns in an urban community: an epidemiological assessment. *Journal of the American Academy of Child & Adolescent Psychiatry* **31**, 951–960.
- ZHAO, J. & SHAO, J. (2014). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association* **accepted**.